Limitations of WER for Intelligibility Evaluation in Speech Anonymization

Victor Ménestrel[1,2]*, Dorothea Kolossa*[1]*, Sebastian Möller*[1]*, Slim Ouni*[2]

[1]Technische Universität Berlin, Germany
[2]Inria, France

{menestrel, dorothea.kolossa, sebastian.moeller}@tu-berlin.de, slim.ouni@loria.fr

Abstract

The increasing deployment of speech technologies in everyday applications has raised significant concerns regarding the protection of personal data and compliance with privacy regulations such as the General Data Protection Regulation (GDPR). Speech signals inherently encode a wide range of biometric and personal attributes beyond linguistic content. These may include speaker identity, age, gender, health status, personality traits, racial or ethnic origin, geographical background, social identity, and socio-economic status [1]. As a result, protecting privacy in speech data has become an active area of research in the field of speech processing.

One promising approach to mitigate privacy leakage risks is speech anonymization, which aims to transform speech signals in such a way that sensitive speaker information cannot be inferred while preserving the linguistic information necessary for downstream tasks. Research in this area has been stimulated by community initiatives such as the Voice Privacy Challenge [2], which provide benchmarks and evaluation frameworks for privacy-preserving speech technologies.

However, anonymization inevitably introduces a fundamental trade-off between privacy and utility. While anonymization techniques attempt to conceal sensitive attributes, speech signals encode multiple paralinguistic cues that may overlap with both privacy-related and task-relevant information. For example, emotional expression can simultaneously convey meaningful communicative content while also revealing personal characteristics [3]. Consequently, the acceptable balance between privacy protection and information preservation depends on the target application and on which speech attributes must remain usable [4].

Reliable evaluation protocols are therefore essential to assess both privacy protection and utility preservation. Previous work has highlighted the risk of overestimating privacy guarantees when evaluation procedures are insufficiently robust [5]. This underlines the importance of ensuring that utility is also evaluated with reliable and well-justified metrics. In the context of speech anonymization, intelligibility is often considered a utility dimension. Most studies estimate intelligibility through the Word Error Rate (WER) produced by automatic speech recognition (ASR) systems, typically evaluated on the test-clean and dev-clean subsets of the LibriSpeech corpus [6].

Although the WER is widely adopted due to its simplicity and reproducibility, it presents several limitations when used as a proxy for intelligibility in anonymized speech. First, the WER was developed to evaluate transcription accuracy in speech-to-text systems, whereas anonymization is primarily a speech-to-speech transformation problem. Consequently, transcription errors measured by WER do not necessarily correspond to intelligibility degradation perceived by human listeners.

Second, qualitative observations on LibriSpeech evaluation sets suggest that a substantial portion of transcription errors arise from proper nouns, spelling variations, or formatting conventions. For example, differences between British and American spellings (e.g., colour vs. color) or numeric forms (42 vs. forty-two) may increase WER despite minimal impact on human comprehension. This raises the question of whether all word-level errors should be treated as equally important for intelligibility evaluation.

Third, WER is inherently system-dependent. Its value depends on the architecture, training data, and decoding strategies of the ASR system used for evaluation. Different ASR systems, such as Whisper [7] or ASR systems provided by toolkits like SpeechBrain [8], may produce different transcription outputs, which could influence WER-based evaluations. As a result, utility estimates based on a single ASR model may not generalize across systems.

These observations motivate several potential research directions for more reliable intelligibility evaluation in speech anonymization. One possibility is the use of a mixture of pretrained ASR models, leveraging diverse architectures and training corpora to reduce model-specific biases. Additionally, incorporating speech intelligibility prediction models commonly used in audiology could provide a valuable complement to ASR-based metrics. Another direction is the use of relative WER, which measures performance degradation relative to the original, non-anonymized speech. Such a metric could help isolate errors specifically introduced by the anonymization process rather than those originating from the dataset or transcription model.

Another aspect is the semantic bias of ASR models used to compute WER. Modern systems rely not only on acoustic signals but also on language models trained on large text corpora, which favor semantically plausible word sequences. While this improves transcription accuracy for natural speech, it may bias the evaluation of anonymized speech by "correcting" acoustically ambiguous signals into meaningful sentences. This raises the question of how ASR systems behave when processing syntactically correct but semantically improbable sentences, where semantic context cannot guide the prediction. Studying such cases on anonymized speech could help disentangle errors caused by acoustic degradation from those compensated by the language model.

Additional complementary metrics may also provide useful insights. For instance, Phoneme Error Rate (PER) could capture pronunciation-level similarities even when orthographic variations differ, although it may be more sensitive to accent changes introduced by anonymization. Moreover, the confidence scores produced by ASR models could potentially serve as indicators of recognition difficulty. Investigating whether such confidence

measures correlate with human-perceived intelligibility remains an open question, particularly since previous studies suggest that intelligibility and perceived speech quality correlate in original speech but not necessarily in anonymized speech [9].

Finally, the reliance on the test-clean subset of LibriSpeech for anonymization evaluation also presents potential limitations. Previous work has shown that speaker identity may still be partially inferred from the linguistic content alone within this dataset [10]. This raises concerns about whether current benchmarks fully isolate speaker identity from speech content, which could bias privacy and utility evaluations. Moreover, current research predominantly focuses on English anonymization using clean, read-aloud speech, while multilingual, noisy, and conversational speech settings remain largely underexplored.

Index Terms: speech anonymization, utility, intelligibility, ASR evaluation.

1. Acknowledgements

This work was funded by the European Union's Horizon Europe research and innovation programme grant No 101168193.

2. References

[1] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, Nov. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0885230818303875

[2] N. Tomashenko, X. Miao, P. Champion, S. Meyer, M. Panariello, X. Wang, N. Evans, E. Vincent, J. Yamagishi, and M. Todisco, "The Third VoicePrivacy Challenge: Preserving Emotional Expressiveness and Linguistic Content in Voice Anonymization," Jan. 2026, arXiv:2601.11846 [cs]. [Online]. Available: http://arxiv.org/abs/2601.11846

[3] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, N. Andrews, and M. Wiesner, "Privacy Versus Emotion Preservation Trade-Offs in Emotion-Preserving Speaker Anonymization," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2024, pp. 409–414. [Online]. Available: https://ieeexplore.ieee.org/document/10832351/

[4] S. Meyer and N. T. Vu, "Use Cases for Voice Anonymization," in *5th Symposium on Security and Privacy in Speech Communication*. ISCA, Aug. 2025, pp. 73–84. [Online]. Available: https://www.isca-archive.org/spsc_2025/meyer25_spsc.html

[5] M. Panariello, S. Meyer, P. Champion, X. Miao, M. Todisco, N. T. Vu, and N. Evans, "The Risks and Detection of Overestimated Privacy Protection in Voice Anonymisation," in *5th Symposium on Security and Privacy in Speech Communication*. ISCA, Aug. 2025, pp. 8–12. [Online]. Available: https://www.isca-archive.org/spsc_2025/panariello25_spsc.html

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210. [Online]. Available: http://ieeexplore.ieee.org/document/7178964/

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[8] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," Jun. 2021, arXiv:2106.04624 [eess]. [Online]. Available: http://arxiv.org/abs/2106.04624

[9] S. T. Arasteh, S. Afza, T.-T. Nguyen, L. Buess, M. Parvin, T. Arias-Vergara, P. A. Perez-Toro, H. C. Hung, M. Lotfinia, T. Gorges, E. Noeth, M. Schuster, S. H. Yang, and A. Maier, "Perceptual Implications of Automatic Anonymization in Pathological Speech," Aug. 2025, arXiv:2505.00409 [eess]. [Online]. Available: http://arxiv.org/abs/2505.00409

[10] C. Franzreb, A. Das, T. Polzehl, and S. Möller, "Content Leakage in LibriSpeech and Its Impact on the Privacy Evaluation of Speaker Anonymization," Jan. 2026, arXiv:2601.13107 [eess]. [Online]. Available: http://arxiv.org/abs/2601.13107